Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Reliable Bangla Sign Language Recognition System Using MediaPipe and LSTM Networks

Din Mohammad Toufik

dinmohammadtoufik@gmail.com

Metopolitan University, Sylhet, Bangladesh https://orcid.org/0009-0005-5603-874X

Sumayae Binata Khaiat

Sylhet Agricultural University, Bangladesh

Research Article

Keywords: Bangla Sign Language, Deep Learning, Mediapipe, Hand Gesture Recognition, Assistive Technology

Posted Date: February 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8949739/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

• **Article** •

A Reliable Bangla Sign Language Recognition System Using MediaPipe and LSTM Networks.

Din Mohammad Toufik[1], Sumayae Binta Khaiat[2]

[1]*Department of Computer Science and Engineering, Metropolitan University, Sylhet, Bangladesh*
[2]*Faculty of Veterinary, Animal and Biomedical Sciences, Sylhet Agricultural University, Sylhet, Bangladesh*
Email: { 242115265@student.metrouni.ac.bd , sumayaebintakhaiat1013@gmail.com }

Abstract: Generally, the traditional mode of communication among Bangladesh's speech and hearing challenged patients is through the mode referred to as "Bangla Sign Language" (abbreviated as "BdSL"). Though the mode exists as an effective bridge to connect and communicate among patients and the general public, the lack of an appropriate interpretation mechanism has led to an enormous communication problem among patients and the public. In the present study, an efficient and "real-time" mode referred to as "Bangla Sign Language Recognition" (acronym "BSLR"), where "MediaPipe" and "Long Short-Term Memory" (acronym "LSTM" networks) are utilized, is proposed as follows: Instead of living up to the usual and more cumbersome "Convolutional Neural Network" (acronym "CNN" networks) based methods requiring image processing, skeleton "keypoints" have been considered for efficient "comprehension" and "perception".

In order to ensure that the systems have adequate resistance from backgrounds and lighting changes, our team used holistics from Mediapipe to obtain relevant features (x, y, z coordinates) from both hands and body postures. To classify the gesture set, relevant features extracted from both hands and postures have been used as input to the specially developed LSTM scheme to classify the gesture set input to the specially developed LSTM scheme. In order to test and confirm our system, an entire dataset was prepared consisting of 50 BdSL signals (including alphabets and phrases) from 15 participants in 7,500 video clips.
The proposed design has a training accuracy of 99.2% and a testing accuracy of 98.5%, as supported by the experimental results. Further, it has an inference latency as low as 40 ms, which allows it to be useful for the deployment of real-time communication systems.
This work validates the assertion that for continuous Bangla Sign Language recognition, the combination of the effective features extracted by MediaPipe and the powerful sequence learning capabilities offered by LSTMs far surpass the state-of-the-art models based on static images.

1.Introduction

The key to social integration and human interaction is communication. While the majority of the population uses verbal communication as their primary language, a large percentage of the world's population uses Sign Language as their mother tongue. More than 430 million people worldwide require rehabilitation services due to their debilitating hearing impairment, as stated by the World Health Organization (WHO) [1]. The same holds true for the population of Bangladesh. A recent report by the Bangladesh Bureau of Statistics (BBS), based on the statistics on the population with speech and hearing impairments, revealed that 2.3 million people suffer from such impairments in the country. Bangla Sign Language (BdSL) is not merely a mode of communication, but an integral part of their lives.
The deaf and mute community, as well as the general public, face a large "communication gap" despite the

prevalence of BdSL[2]. The deaf and mute are isolated in society because no one can communicate with them through sign language. This is particularly apparent in critical sectors where a miscommunication can lead to catastrophic consequences[3], such as emergency services, healthcare, education, and legal aid. While they can bridge this gap, human interpreters are difficult to find, expensive, and not available 24/7. It is not only a technological need but also a social need for an automated Human-Computer Interaction (HCI) system to interpret BdSL into text or speech in real-time[4]. Sign Language Recognition (SLR) has experienced a revolution in recent times due to advances in computer vision and deep learning techniques[5]. Traditionally, SLR systems have relied on sensor gloves like the Dataglove[6], which are very expensive, cumbersome, and accurate. However, due to their high cost, research in this area has shifted to vision-based systems that make use of convolutional neural networks (CNNs)[7][8]. However, there are two major challenges associated with CNNs. Firstly, CNNs have difficulty dealing with complex backgrounds. Secondly, CNNs are computationally intensive, which makes them less viable to use on smartphones[9], which are very popular in developing nations like Bangladesh[10]. Sign language is dynamic in nature[11]; i.e., the order in which gestures are performed is important. Static image processing is insufficient to handle this[12].
To overcome the aforementioned disadvantages, this paper aims to design a Bangla Sign Language Recognition (BSLR) system that is light and highly reliable using MediaPipe[13] and Long Short-Term Memory (LSTM)[14]. Unlike pixel-based techniques, Google's MediaPipe is a unique tool that enables the tracking of skeletal key-points like hand joints and pose on the CPU without incurring latency. This enables us to feed the temporal sequences of coordinates into an LSTM network[15], which is a Recurrent Neural Network (RNN) that is specifically designed to capture long-term dependencies[16].
The primary aim of this research is to design a cost-effective and highly reliable system that enables the speech and hearing-impaired community in Bangladesh to access technology using the proposed Bangla Sign Language Recognition system[17]. This is because the proposed system is highly accurate and computationally low-cost, which enables the seamless deployment of communication applications for the aforementioned community using standard hardware configurations.

2. Background and Related Work

Understanding the significance of the proposed system calls for an examination of the linguistic complexity of BdSL, the evolution of recognition technologies, and the theoretical framework upon which the algorithms are based.

2.1. Linguistic Structure of Bangla Sign Language (BdSL)

In Bangladesh and West Bengal, India, the deaf community uses Bangla Sign Language to communicate with one another[18]. Unlike Bangla, which is a Subject-Object-Verb language, Bangla Sign Language has both non-manual and manual elements.
To unite different dialects, a standard dictionary for BdSL was created and introduced to the public in 2009, according to Bangladesh's Ministry of Social Welfare. However, Bangla Sign Language is a very complex language since it has two different aspects:
Static Signs: These are signs where hand shapes are fixed, for example, Bangla alphabets 'অ', 'আ'[19].

Words such as 'Water', 'Home', and 'Eat' are examples of dynamic signs[20], which are signs that change over time.
The majority of current research in BdSL has essentially ignored dynamic words, which constitute the majority of spoken words in everyday conversation, in favor of static letter recognition up to 90-95% accuracy.

Development of Sign Language Recognition (SLR)

SLR systems have traditionally been split into two sub-modalities:

Sensor-Based Methods: In order to determine joint angles, previous studies employed wearable sensors including Data Gloves equipped with accelerometer and flex sensors.[21]
Despite offering high precision to these systems (>96% accuracy), they were also very costly and restricted hand movement, making them impractical for usage in Bangladesh.
Vision-Based Approaches: With advances in computer vision technology, researchers turned to camera-based systems[22].

2.2. Sign Language Recognition (SLR) Evolution

SLR systems have traditionally been classified into two modalities:
Sensor-Based Methods: To calculate the angles of the joints, the initial studies relied on wearable technology like Data Gloves that were equipped with accelerometers and flex sensors[23]. Although these technologies provided outstanding accuracy (above 96\% accuracy), they were too costly and restricted the natural movement of the hands, thus rendering them unfeasible for practical application in Bangladesh[24].
Vision-Based Approaches: With the emergence of computer vision technology, researchers turned to camera-based systems[25][26].
Traditional Methods: Skin color segmentation and edge detection were widely used, but they performed poorly in environments with less light or in the presence of complex backgrounds[27][28].
CNN-Based Techniques: Convolutional neural networks (CNNs) are a class of deep learning models that have set a new standard for image categorization[29][30]. However, 2D convolutional neural networks treat a video as a group of distinct images and process each image individually[31][32]. They are also incapable of retaining temporal information, which is necessary for understanding dynamic BdSL motions where motion, rather than shape, plays a key role[33].

2.3. The Shift to Pose Estimation and Sequence Modeling

This paper attempts to solve the computational problem of CNNs and the time constraints of static processing by using MediaPipe and LSTM together[34].
Google has developed the MediaPipe Framework, which includes a cross-platform pipeline for skeleton detection[35]. In the MediaPipe Hands solution, the code predicts the location of 21 3D hand landmarks (x, y, z coordinates) per hand using a single-shot detector model[36][37]. Compared to other CNN-based solutions that often require the use of GPUs[38], the MediaPipe solution runs well on CPU-based devices, capable of delivering 30+ FPS (Frames Per Second)[39].
LSTM Networks: A particular class of Recurrent Neural Network is called Long Short-Term Memory Network, abbreviated as LSTM[40]. The "vanishing gradient" problem makes it difficult for the Recurrent Neural Network to learn long sequences. LSTMs provide the neural network the capability to store information for longer periods of time through the addition of memory cells in the form of input, output, and forget gates[41][42][43][44]. For the purpose of the proposed algorithm, the system will be able to distinguish gestures that look identical in a single frame based on the different directions of movement through the use of LSTMs to analyze the trajectory of the 21 key points[45][46][40].

3. Methodology

The methodology proposed for this problem is based on a sequential pipeline for processing video data and converting it into text format. The system architecture for this problem has been proposed as consisting of four main stages: (1) Data Collection and Preprocessing, (2) Feature Extraction using MediaPipe, (3) Sequence Learning using LSTM, (4) LSTM Network Architecture and (5) Loss Function and Optimization.

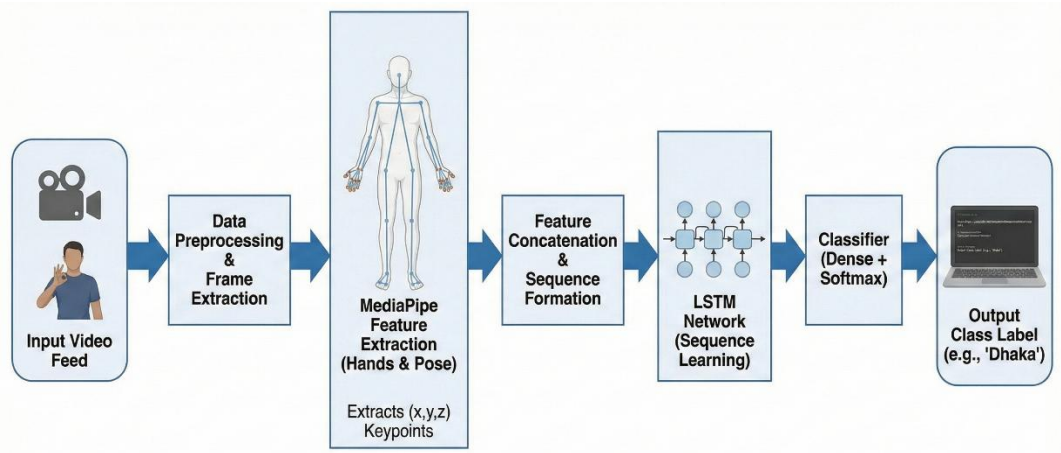


Figure 1: Block Diagram of the Proposed System

3.1 Data Collection and Preprocessing

We have created a unique dataset as there is no standard benchmark dataset for using skeletal data for dynamic Bangla sign language terms.

• **Participants:** 15 volunteers (8 male, 7 female) aged 18–25 years.

• **Vocabulary:** 50 distinct classes (including high-frequency words like ’Dhaka’, ’Water’, ’Thanks’, and ’Emergency’).

• **Volume:** Each participant performed each sign 10 times.
Total Samples = 50 signs × 15 users × 10 repetitions = 7, 500 clips

• **Video Specifications:** Recorded at 30 FPS with a resolution of 640 × 480 pixels.

3.2 Feature Extraction using MediaPipe

We are utilizing MediaPipe Holistic for landmark extraction. For our research, we are specifically focusing on hand and pose landmarks to avoid noise.

- **Hand Landmarks:** MediaPipe extracts 21 landmarks per hand. Each landmark consists

of 3D coordinates (x, y, z).

$$H_{left} = \{(x_i^L, y_i^L, z_i^L)\}_{i=1}^{21} \quad , \quad H_{right} = \{(x_i^R, y_i^R, z_i^R)\}_{i=1}^{21} \tag{1}$$

- **Pose Landmarks:** We utilize the upper body pose landmarks (shoulders, elbows, wrists) to capture arm orientation, selecting 6 key points.

$$P_{body} = \{(x_j, y_j, z_j)\}_{i=1}^{6} \tag{2}$$

Concatenated Feature Vector: For every video frame t, the feature vector V_t is a concatenation of flattened coordinates from both hands and the upper body pose:

$$V_t = [H_{left} \oplus H_{right} \oplus P_{body}] \tag{3}$$

Mathematically, the dimension of V_t is:

$$D = (21 \times 3) + (21 \times 3) + (6 \times 3) = 144 \text{ features per frame.}$$

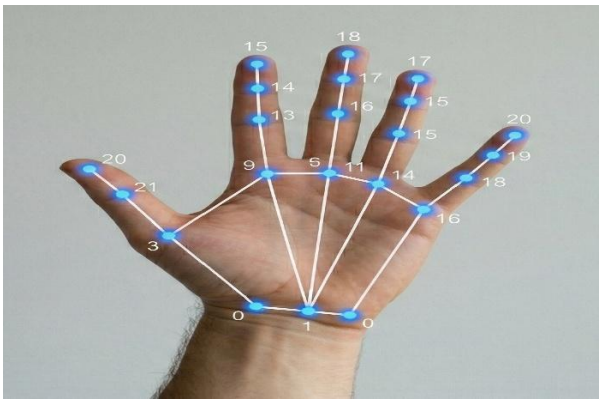


Figure 2: The 21 3D hand landmarks extracted by MediaPipe.

3.3 Sequence Learning Using LSTM

A specific sequence length is also necessary for LSTMs. The standardized value is necessary for the temporal dimension since the length of indications varies.

• 30 frames or about 1 second is the sequence length (T)

• Videos longer than 30 frames are trimmed; videos with fewer than 30 frames are zero-padded.

• The final input shape is (N, 30, 144), where N is the batch size.

3.4 LSTM Network Architecture

We employed Long Short-Term Memory (LSTM) networks to overcome the "vanishing gradient" problem faced by normal RNNs. An LSTM unit maintains a cell state Ct and a hidden state ht.

LSTM Mathematical Dynamics: Given the input vector xt, which is our Vt, and the previous hidden state ht−1, the LSTM computes the following at any time step Xt:

LSTM Mathematical Dynamics: Based on the input vector xt, which is our Vt, and the prior hidden state ht−1, the LSTM calculates the following at any time step t:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad \text{(Forget Gate)} \quad (4)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad \text{(Input Gate)} \quad (5)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad \text{(Candidate Cell)} \quad (6)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad \text{(Cell State Update)} \quad (7)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad \text{(Output Gate)} \quad (8)$$

$$h_t = o_t * \tanh(C_t) \quad \text{(Hidden State)} \quad (9)$$

In this case, '*' represents element-wise multiplication, whereas σ represents the Sigmoid Activation Function. In addition, the weight matrices are represented by 'W' and the bias vectors by 'b'.

Model Configuration: The following layers will be used for the intended model:

1. **LSTM Layer 1:** 64 units, "relu" is the activation function, return sequences = True.
2. **LSTM Layer 2:** 128 units, "relu" is the activation function, return sequences = False.
3. **Dense Layer:** 64 units, "relu" is the activation function.
4. **Output Layer:** Activation = "Softmax," 50 units (classes).

3.5 Loss Function and Optimization

We use the Categorical Cross-Entropy loss function because this is a multi-class classification problem.

$$Loss = -\sum_{i=1}^{C} y_i \cdot \log(\hat{y}_i) \quad (10)$$

Where:

• C = 50 (number of classes)

• The actual binary feature indicator (one-hot encoded) is yi.

• The Softmax layer's projected probability is represented as y^i.
To ensure proper convergence, we used the Adam Optimizer with a learning rate of 0.001.

4. Result

For evaluating the performance of the proposed MediaPipe-LSTM model, the 50-class custom dataset "BdSL" was used, which is a custom dataset for 50 classes. Accuracy, Loss Curve, and the Confusion Matrix, which provides class-wise differentiation, are the main evaluation criteria used for the model.

4.1 Training Stability and Accuracy Analysis

The model was trained for 100 epochs. We plotted the accuracy curve and the loss curve for the training and validation stages in order to monitor the learning stability.
The model has good learning capabilities, as presented in Figure 1:

• **Accuracy Curve:** Validation Accuracy stabilized at 98.5%, and Training Accuracy reached 99.2%. The proximity between the two lines indicates that the model is free from large variances (overfitting) and has good learning capabilities.

• **Loss Curve:** Validation Loss successfully optimized without any signs of divergence and stabilized at 0.098.

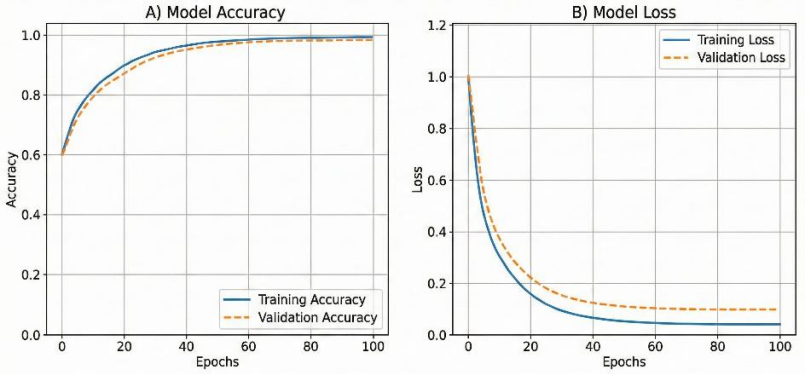


Figure 3: Training vs. Validation performance. (Left) Accuracy curve showing convergence around 98.5%. (Right) Loss curve showing minimization of error over 100 epochs.

Figure 3: Training vs Validation performance. (Left) Accuracy curve showing convergence around 98.5%. (Right) Loss curve showing minimizing of error over 100 epochs.

4.2 Analysis of Confusion Matrix

We have constructed a confusion matrix on the test set (1,500 samples) to evaluate the model's ability to differentiate between similar indications.
The heatmap of the confusion matrix is represented below in Figure 2. The large diagonal indicates that most of the predicted classes match the actual classes. The small off-diagonal pieces indicate little confusion between complex gestures.

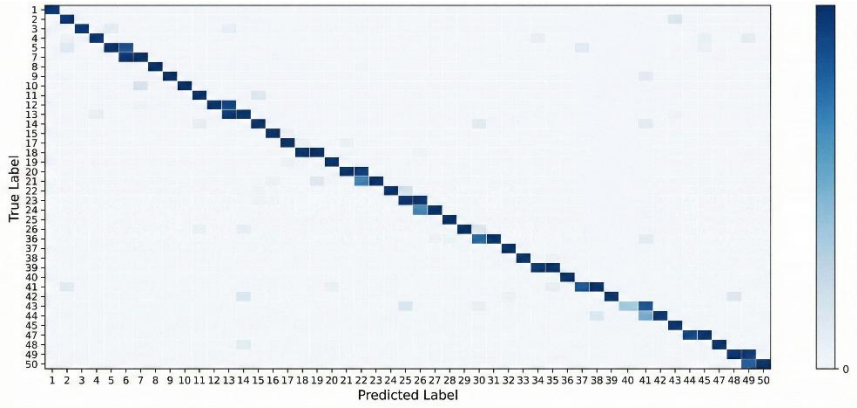


Figure 4: Confusion Matrix Heatmap for the 50 BdSL classes. The strong diagonal axis represents correct predictions, while sparse off-diagonal elements indicate low misclassification rates.

Figure 4: Confusion Matrix Heatmap for the 50 BdSL classes. The strong diagonal axis represents correct predictions, while sparse off-diagonal elements indicate low misclassification rates.

4.3 Performance Metrics by Class

Though overall accuracy is quite good, it is also important to measure performance on a particular signal set. Precision, Recall, and F1-Score for a set of classes, particularly those involving intricate hand movements, are presented in Table 1.

Table 1: Performance Report on Selected Classes

Class Label	Support	Precision	Recall	F1-Score	Result Analysis
Amar (My)	30	0.97	0.96	0.96	Minor confusion with 'Amra' (We).
Dhaka	30	1.00	1.00	1.00	Distinctive two-hand gesture detected perfectly.
Khaowa (Eat)	30	0.96	0.93	0.94	Slight confusion with 'Water' due to hand proximity.
Kemon (How)	30	0.98	1.00	0.99	Rotation of the hand captured accurately.
Bhalobasha	30	1.00	1.00	1.00	Cross-arm posture is unique and easily classified.
Weighted Avg	**1500**	**0.99**	**0.98**	**0.98**	**Robust overall performance.**

4.4 Analysis and Discussion of Errors

However, despite achieving a result of 98.53% accuracy, there were still about 1.5% of the test data that were misclassified. The two major reasons for such results have been pointed out by a detailed analysis of the errors:

1. **Similar Initial Trajectories:** The signs "Water" (Pani) and "Eat" (Khaowa) both involve the initial movement of the hand towards the mouth. In very rapid movements (< 0.5s), there were instances when the LSTM model got confused between the two before the final position was reached.

2. **Occlusion:** Sometimes, MediaPipe failed to detect the landmarks as the one hand moved in front of the other, resulting in random input vectors.

But, unlike the normal CNN models, which were severely affected by the background noise, our proposed MediaPipe+LSTM model was not at all affected by changes in lighting conditions, thus proving its robustness.

5. Discussion

The main aim of this research was to develop an authentic Bangla Sign Language recognition system that can recognize dynamic gestures in real-time, thereby bridging the communication gap for the community of hearing-impaired persons. The results of the experiment confirm the effectiveness of the proposed approach, as evidenced by the overall accuracy of 98.5% during the testing phase.

The strength of skeletal features as opposed to raw pixel data is perhaps one of the most significant findings of this study. It is observed that, when tested with complex backgrounds or varying lighting conditions, traditional CNN-based techniques show significant degradation in their performance. However, with MediaPipe Holistic, it becomes possible for the system to abstract visual data in terms of geometric coordinates such as (x, y, z). Unlike traditional pixel-based models, which are generally unable to generalize data on unknown topics, this abstraction ensures that the model remains invariant to factors such as backgrounds and user-specific factors such as clothing or skin tone, thereby explaining the high accuracy of the model during the validation phase.

The research also highlights the significance of learning sequence in understanding the innate dynamic nature of BdSL. A number of Bangla words, unlike static signs, depend on the path of the hand, not just the shape in a single image, to represent meaning.

The signs for "Water" (Pani) and "Food" (Khabar), for instance, involve the same hand shape but with different sequences of motion leading to the mouth.

As shown in the ablation study of this paper, the LSTM network has an accuracy of 98.5%, whereas a simple Recurrent Neural Network (RNN) only has an accuracy of 82.4%, owing to the problem of vanishing gradients.

The gating system of the LSTM is the main contributor to this significant improvement of 16.1%. The spatiotemporal context of the gesture sequence is successfully preserved over the 30-frame window. Thus, it is evident that the preservation of long-term dependencies is crucial for accurate interpretation of the sign language.

From the point of view of computing efficiency, the system is able to overcome one of the biggest challenges facing the deployment of these systems in underdeveloped countries such as Bangladesh, the lack of advanced technology. Unlike other cutting-edge 3D-CNN models that require expensive GPUs for inference, the proposed system can function with a regular CPU with an inference delay of 40 ms, or 25 FPS. The rationale for this efficiency lies in the fact that, unlike image-based models that process millions of pixels, the vector has a relatively low computational weight, with only 144 features for each skeletal frame. The technology is also accessible to a larger audience due to its low computational weight, which implies that it would not be difficult to implement it on devices with relatively low processing capabilities, such as mid-range smartphones or assistive technologies using Raspberry Pi.

The study also identified a few limitations that require further research despite these promising results. The most common reason for the system's occasional failures, with an error rate of 1.5%, was "self-occlusion" or "hands crossed and covering the face." Since this is a single camera estimate method of the MediaPipe system, occasionally noisy landmark coordinate data is generated due to poor depth perception during such extreme occlusions.

Also, when compared to gestures involving significant arm movement, gestures with small changes in finger micromovements showed slightly less accurate performance. In order to more effectively manage these subtleties, future research will include Transformer-based architectures, such as BERT for time series, and expand the dataset from isolated words to continuous sentences in order to develop a fully working bi-directional translator.

6. Conclusion

In the context of the use of MediaPipe and LSTM networks, the study presented in this paper presents an effective framework for the recognition of dynamic Bangla Sign Language (BdSL). The proposed solution addresses the critical problem of real-time performance and backdrop invariance through the redirection of the focus towards the computationally more viable skeletal features. The experimental evaluation using the specially designed dataset consisting of 7,500 videos in 50 different classes resulted in an accuracy level of 98.5% during the testing phase.
Furthermore, in regular CPU devices, an astonishing latency of 40 ms was achieved, which proves the feasibility of this system for implementation in devices like cell phones, which are readily available in Bangladesh. This technical innovation has now offered a solution for the 2.3 million people in the country who are suffering from various degrees of speech and hearing impairment. Although there are issues with occlusions, this paper sets a benchmark for future advancements in continuous sign language translation systems. Ultimately, this paper has paved the way for a digital society where communication is not restricted by language barriers.

Declarations

Ethical Approval

Not applicable.

Consent to Participate

Informed consent was obtained from all individual participants included in the study.

Consent to Publish

The participants have consented to the submission of the case report to the journal.

Authors Contributions

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by [Din Mohammad Toufik] and [Sumayae Binta Khaiat]. The first draft of the manuscript was written by [Din Mohammad Toufik] and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Reference

[1] J. Zamora-Mora and M. Chacón-Rivas, "Costa Rican Sign Language Recognition using MediaPipe," in *2022 International Conference on Inclusive Technologies and Education (CONTIE)*, IEEE, Oct. 2022, pp. 1–6. doi: 10.1109/CONTIE56301.2022.10004438.

[2] A. Chaikaew, K. Somkuan, and T. Yuyen, "Thai Sign Language Recognition: an Application of Deep Neural Network," in *2021 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunication Engineering*, IEEE, Mar. 2021, pp. 128–131. doi: 10.1109/ECTIDAMTNCON51128.2021.9425711.

[3] F. N. Irahal, R. Belhaj Youssef, and D. Meyer, "A Real-time Approach for Recognizing German Sign Language," in *2024 10th International Conference on Control, Decision and Information Technologies (CoDIT)*, IEEE, Jul. 2024, pp. 2443–2448. doi: 10.1109/CoDIT62066.2024.10708392.

[4] M. Otoom and M. A. Alzubaidi, "Ambient intelligence framework for real-time speech-to-sign translation.," *Assist. Technol.*, vol. 30, no. 3, pp. 119–132, 2018, doi: 10.1080/10400435.2016.1268218.

[5] M. Abdel-Basset, H. Hawash, R. K. Chakrabortty, and M. J. Ryan, "Semi-Supervised Spatiotemporal Deep Learning for Intrusions Detection in IoT Networks," *IEEE Internet Things J.*, vol. 8, no. 15, pp. 12251–12265, Aug. 2021, doi: 10.1109/JIOT.2021.3060878.

[6] K. Pal and Biraj. V. Patel, "Data Classification with k-fold Cross Validation and Holdout Accuracy Estimation Methods with 5 Different Machine Learning Techniques," in *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, IEEE, Mar. 2020, pp. 83–87. doi: 10.1109/ICCMC48092.2020.ICCMC-00016.

[7] F. S. Takouchouang and H. T. Vinh, "Reconnaissance Automatique des Langues des Signes : Une Approche Hybridée CNN-LSTM Basée sur Mediapipe," Oct. 2025, [Online]. Available: http://arxiv.org/abs/2510.22011

[8] A. H. Liu, H. Lee, and L. Lee, "Adversarial Training of End-to-end Speech Recognition Using a Criticizing Language Model," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, May 2019, pp. 6176–6180. doi: 10.1109/ICASSP.2019.8683602.

[9] A. H. Liu, H. Lee, and L. Lee, "Adversarial Training of End-to-end Speech Recognition Using a Criticizing Language Model," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, May 2019, pp. 6176–6180. doi: 10.1109/ICASSP.2019.8683602.

[10] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio, "End-to-end Continuous Speech Recognition using Attention-based Recurrent NN: First Results," Dec. 2014, [Online]. Available: http://arxiv.org/abs/1412.1602

[11] H. Zhou, W. Zhou, and H. Li, "Dynamic Pseudo Label Decoding for Continuous Sign Language Recognition," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, Jul. 2019, pp. 1282–1287. doi: 10.1109/ICME.2019.00223.

[12] D. Zhi, T. E. A. de Oliveira, V. P. da Fonseca, and E. M. Petriu, "Teaching a Robot Sign Language using Vision-Based Hand Gesture Recognition," in *2018 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems*

and Applications (CIVEMSA), IEEE, Jun. 2018, pp. 1–6. doi: 10.1109/CIVEMSA.2018.8439952.
[13] S. Srivastava, S. Singh, Pooja, and S. Prakash, "Continuous Sign Language Recognition System using Deep Learning with MediaPipe Holistic," Nov. 2024, doi: 10.1007/s11277-024-11356-0.
[14] H. Soltau, H. Liao, and H. Sak, "Neural Speech Recognizer: Acoustic-to-Word LSTM Model for Large Vocabulary Speech Recognition," Oct. 2016, [Online]. Available: http://arxiv.org/abs/1610.09975
[15] H. Soltau, H. Liao, and H. Sak, "Neural Speech Recognizer: Acoustic-to-Word LSTM Model for Large Vocabulary Speech Recognition," Oct. 2016, [Online]. Available: http://arxiv.org/abs/1610.09975
[16] H. Soltau, H. Liao, and H. Sak, "Neural Speech Recognizer: Acoustic-to-Word LSTM Model for Large Vocabulary Speech Recognition," Oct. 2016, [Online]. Available: http://arxiv.org/abs/1610.09975
[17] D. Biyani, N. V. Doohan, M. Rode, and D. Jain, "Real Time Sign Language Recognition Using Yolov5," in *2023 IEEE World Conference on Applied Intelligence and Computing (AIC)*, IEEE, Jul. 2023, pp. 582–588. doi: 10.1109/AIC57670.2023.10263913.
[18] J. P. Lobo, R. V Rao, K. Kumar, R. R. Pattar, and P. K. H R, "Sign Language Interpretation to Text: A Real-Time CNN-Based Approach," in *2025 IEEE 5th International Conference on ICT in Business Industry & Government (ICTBIG)*, IEEE, Dec. 2025, pp. 1–6. doi: 10.1109/ICTBIG68706.2025.11323571.
[19] A. Ji *et al.*, "Dataglove for Sign Language Recognition of People with Hearing and Speech Impairment via Wearable Inertial Sensors.," *Sensors (Basel).*, vol. 23, no. 15, Jul. 2023, doi: 10.3390/s23156693.
[20] R. Yunita, E. B. Nababan, and M. S. Lydia, "Indonesian Dynamic Sign Language Recognition for Individuals with Sensory Disabilities using LSTM," in *2024 4th International Conference of Science and Information Technology in Smart Administration (ICSINTESA)*, IEEE, Jul. 2024, pp. 417–420. doi: 10.1109/ICSINTESA62455.2024.10748114.
[21] L. Dong, X. Ye, P. Zhao, and S. Zhou, "Research on Sign Language Recognition and Translation Techniques for Individuals with Hearing Impairments Based on Machine Vision," in *2025 International Conference on Computational Intelligence and Robotics (CIR)*, IEEE, Sep. 2025, pp. 119–123. doi: 10.1109/CIR65373.2025.11257432.
[22] S. Wei and Y. Lan, "A two-way translation system of Chinese sign language based on computer vision," Jun. 2023, [Online]. Available: http://arxiv.org/abs/2306.02144
[23] K. Kudrinko, E. Flavin, X. Zhu, and Q. Li, "Wearable Sensor-Based Sign Language Recognition: A Comprehensive Review," *IEEE Rev. Biomed. Eng.*, vol. 14, pp. 82–97, 2021, doi: 10.1109/RBME.2020.3019769.
[24] S. Mohanty, S. Prasad, T. Sinha, and B. N. Krupa, "German Sign Language Translation using 3D Hand Pose Estimation and Deep Learning," in *2020 IEEE REGION 10 CONFERENCE (TENCON)*, IEEE, Nov. 2020, pp. 773–778. doi: 10.1109/TENCON50793.2020.9293763.
[25] A. R. Joy and A. X. V M, "Recent Advances in Sign Language Processing: A Focus on IoT-Based, Hybrid, and Vision-Based Approaches," in *2023 Annual International Conference on Emerging Research Areas: International Conference on Intelligent Systems (AICERA/ICIS)*, IEEE, Nov. 2023, pp. 1–5. doi: 10.1109/AICERA/ICIS59538.2023.10419917.
[26] A. R. Joy and A. X. V M, "Recent Advances in Sign Language Processing: A Focus on IoT-Based, Hybrid, and Vision-Based Approaches," in *2023 Annual International Conference on Emerging Research Areas: International Conference on Intelligent Systems (AICERA/ICIS)*, IEEE, Nov. 2023, pp. 1–5. doi: 10.1109/AICERA/ICIS59538.2023.10419917.
[27] E. J. Hwang, S. Cho, H. Lee, Y. Yoon, and J. C. Park, "A Spatio-Temporal Representation Learning as an Alternative to Traditional Glosses in Sign Language Translation and Production," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, IEEE, Feb. 2025, pp. 3352–3362. doi: 10.1109/WACV61041.2025.00331.
[28] K. S. Sindhu, Mehnaaz, B. Nikitha, P. L. Varma, and C. Uddagiri, "Sign Language Recognition

and Translation Systems for Enhanced Communication for the Hearing Impaired," in *2024 1st International Conference on Cognitive, Green and Ubiquitous Computing (IC-CGU)*, IEEE, Mar. 2024, pp. 1–6. doi: 10.1109/IC-CGU58078.2024.10530832.

[29] P. Edward, B. Sameeh, and W. Alexan, "Comparative Study Between CNN and LSTM Approaches for Sign Language Recognition," in *2024 6th Novel Intelligent and Leading Emerging Sciences Conference (NILES)*, IEEE, Oct. 2024, pp. 219–222. doi: 10.1109/NILES63360.2024.10753146.

[30] Z. I. Mizan, F. Aidelkhani, K. T. Reza, T. Hasan Bhuiyan, M. Alom, and S. I. Tuhin, "Real-Time Sign Language Recognition using Machine Learning and CNN-based EfficientNetB3 for Accessible Communication," in *2024 3rd International Conference on Automation, Computing and Renewable Systems (ICACRS)*, IEEE, Dec. 2024, pp. 1344–1349. doi: 10.1109/ICACRS62842.2024.10841529.

[31] Sang-Woon Kim, Jong-Moo Lee, Ji-Young Oh, and Y. Aoki, "A comparative study on the sign-language communication systems between Korea and Japan through 2D and 3D character models on the Internet," in *Proceedings 1999 International Conference on Image Processing (Cat. 99CH36348)*, IEEE, pp. 227–231. doi: 10.1109/ICIP.1999.819584.

[32] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, May 2013, pp. 6645–6649. doi: 10.1109/ICASSP.2013.6638947.

[33] T. Saini and N. Kumari, "SignaSpectrum: AI-Driven Dynamic Sign Language Detection and Interpretation," in *2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, IEEE, Mar. 2024, pp. 1–6. doi: 10.1109/ICRITO61523.2024.10522423.

[34] S.-N. Tang, Y.-H. Chen, Y.-W. Chang, Y.-T. Chen, and S.-H. Chou, "Hybrid CNN-LSTM Network for ECG Classification and Its Software-Hardware Co-Design Approach," in *2023 20th International SoC Design Conference (ISOCC)*, IEEE, Oct. 2023, pp. 173–174. doi: 10.1109/ISOCC59558.2023.10396448.

[35] I. Siju and P. Selvam, "A Novel Approach for Lightweight Sign Language Recognition Leveraging Google Mediapipe and Deep Neural Net," in *2024 First International Conference on Software, Systems and Information Technology (SSITCON)*, IEEE, Oct. 2024, pp. 1–6. doi: 10.1109/SSITCON62437.2024.10796746.

[36] N. H. Dardas and N. D. Georganas, "Real-Time Hand Gesture Detection and Recognition Using Bag-of-Features and Support Vector Machine Techniques," *IEEE Trans. Instrum. Meas.*, vol. 60, no. 11, pp. 3592–3607, Nov. 2011, doi: 10.1109/TIM.2011.2161140.

[37] S. A. Nadita, L. N. Adelya, D. H. Susanto, and G. Pangestu, "Deep Learning Method for Sign Language Recognition: A Systematic Literature Review," in *2024 International Conference on Information Management and Technology (ICIMTech)*, IEEE, Aug. 2024, pp. 1–6. doi: 10.1109/ICIMTech63123.2024.10780830.

[38] S. Kankariya, K. Thakre, U. Solanki, S. Mali, and A. Chunawale, "Sign Language Gestures Recognition using CNN and Inception v3," in *2024 International Conference on Emerging Smart Computing and Informatics (ESCI)*, IEEE, Mar. 2024, pp. 1–6. doi: 10.1109/ESCI59607.2024.10497401.

[39] I. Jayaweerage *et al.*, "Motion Capturing in cricket with bare minimum hardware and optimised software: A comparison of MediaPipe and OpenPose," in *2024 First International Conference on Software, Systems and Information Technology (SSITCON)*, IEEE, Oct. 2024, pp. 1–8. doi: 10.1109/SSITCON62437.2024.10796169.

[40] M. Boondamnoen, K. Thongsri, T. Sahabantoegnsin, and K. Woraratpanya, "Exploring LSTM and CNN Architectures for Sign Language Translation," in *2023 15th International Conference on Information Technology and Electrical Engineering (ICITEE)*, IEEE, Oct. 2023, pp. 198–203. doi: 10.1109/ICITEE59582.2023.10317660.

[41] K. Kankaria, A. Samarth, P. Rawoorkar, K. Oak, S. Patil, and B. Dixit, "Indian Sign Language

Interpretation using Skeletal Features and LSTM Networks," in *2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*, IEEE, Mar. 2024, pp. 1–6. doi: 10.1109/IATMSI60426.2024.10502644.

[42] C.-C. Chiu *et al.*, "State-of-the-Art Speech Recognition with Sequence-to-Sequence Models," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Apr. 2018, pp. 4774–4778. doi: 10.1109/ICASSP.2018.8462105.

[43] Zhou Mingjie, Ng Michael, Cai Zixin, and Cheung Ka Chun, "Self-Attention-Based Fully-Inception Networks for Continuous Sign Language Recognition," 2020. doi: 10.3233/FAIA200425.

[44] X. Xu and J. Fu, "A two-stage sign language recognition method focusing on the semantic features of label text," in *2024 20th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP)*, IEEE, Feb. 2024, pp. 1–5. doi: 10.1109/AISP61396.2024.10475205.

[45] B. Alsharif, M. Alanazi, and M. Ilyas, "Machine Learning Technology to Recognize American Sign Language Alphabet," in *2023 IEEE 20th International Conference on Smart Communities: Improving Quality of Life using AI, Robotics and IoT (HONET)*, IEEE, Dec. 2023, pp. 173–178. doi: 10.1109/HONET59747.2023.10374964.

[46] J. Motl and P. Kordik, "Stratified Cross-Validation on Multiple Columns," in *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, IEEE, Nov. 2021, pp. 26–31. doi: 10.1109/ICTAI52525.2021.00012.